Geophysical Journal International

Geophys. J. Int. (2014) **198,** 1419–1430 doi: 10.1093/gji/ggu201
GJI Seismology

Bayesian historical earthquake relocation: an example from the 1909 Taipei earthquake

Sarah E. Minson[1] and William H. K. Lee[2]

[1]*Seismological Laboratory, Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, CA* 91125, *USA. E-mail: minson@gps.caltech.edu*
[2]*U.S. Geological Survey, Menlo Park, CA* 94025, *USA*

Accepted 2014 May 28. Received 2014 May 27; in original form 2014 February 18

SUMMARY

Locating earthquakes from the beginning of the modern instrumental period is complicated by the fact that there are few good-quality seismograms and what traveltimes do exist may be corrupted by both large phase-pick errors and clock errors. Here, we outline a Bayesian approach to simultaneous inference of not only the hypocentre location but also the clock errors at each station and the origin time of the earthquake. This methodology improves the solution for the source location and also provides an uncertainty analysis on all of the parameters included in the inversion. As an example, we applied this Bayesian approach to the well-studied 1909 M_w 7 Taipei earthquake. While our epicentre location and origin time for the 1909 Taipei earthquake are consistent with earlier studies, our focal depth is significantly shallower suggesting a higher seismic hazard to the populous Taipei metropolitan area than previously supposed.

Key words: Earthquake dynamics; Earthquake ground motions; Earthquake source observations; Seismicity and tectonics; Site effects.

1 INTRODUCTION

Given the fact that the repeat time of great earthquakes can be hundreds of years or more, it is important to be able to determine earthquake locations for as long a time frame as possible in order to ascertain the seismic hazard of a region. The use of pendulum seismometres began to spread in the late 19th century, but earthquake information from these early days is often limited to little more than *P*- and *S*-wave phase picks at a small number of stations distributed around the globe. These phase picks can contain large errors since teleseismic waveforms often appear to be low amplitude and very emergent on seismograms recorded before the introduction of high-gain seismometres in the 1950s. Even worse, the times of these phase picks are based on the local clock at each seismographic station, and these clocks were often inaccurate. In general, given the lack of sufficient data, the large errors on the available arrival times, and the non-linearity of location inversion using traveltime data, historical hypocentre location problems are commonly numerically unstable and poorly constrained.

One of the first methods developed to compute the locations of earthquakes based on observed traveltimes of seismic phases is to linearize the non-linear equations describing traveltime as a function of hypocentre location (e.g. Geiger 1910; Bolt 1960; Engdahl & Gunst 1966; Lee & Stewart 1981; Thurber 2011). However, the non-linear earthquake location problem can be solved directly through direct search or Monte Carlo simulation, with several authors presenting methods for using Bayesian inference and maximum likelihood estimates to determine earthquake hypocentres (e.g. Jordan & Sverdrup 1981; Tarantola & Valette 1982; Lomax 2005; Lomax *et al.* 2009). Mosegaard & Tarantola (2002) present a useful review of the probabilistic approach to inverse problems that expands on Tarantola & Valette (1982) and includes more recent developments in the field. Most recently, several studies have examined using Bayesian analysis of huge modern data sets to better solve the earthquake location problem for present-day events (e.g. Myers *et al.* 2007, 2011).

Historical earthquakes, for which the recorded phase arrival times are often shifted due to clock errors, are most commonly located by inverting the *S–P* time (the time interval between the *P*- and *S*-wave arrival). This has the advantage of removing the effects of any clock errors because the clock error will shift the recorded time of the *P*- and *S*-wave arrivals equally and thus be canceled by differencing the two times. However, to compute the earthquake origin time, the absolute arrival time of at least one seismic phase is required; *S–P* times alone are not sufficient. Thus, the origin time is often computed by a separate analysis of arrival times at a few seismographic observatories which have a reputation for accurate timekeeping. However, relying on absolute times from only a few stations can lead to biased solutions if any of these stations' clocks or seismic phase arrival picks are inaccurate.

In this paper, we suggest an earthquake location approach that uses Bayesian inference to solve for the hypocentre location using the observed seismic phase arrival times (instead of relative *S–P* time) while simultaneously solving for the origin time of the

earthquake and the clock error at each seismographic station. Our approach has the advantage that we can use all available observations to constrain not only the earthquake focus but also the origin time (similar to Tarantola & Valette 1982), and we can also recover the clock error at each station allowing a retrospective analysis of the reliability of each seismographic observatory. Although solving for a clock error at each observatory yields an inverse problem that formally has almost as many unknowns as observations (if only *P*- and *S*-wave arrival times are available), our formulation of the inverse problem utilizes an analytical solution for the origin time and clock errors so that we can compute the posterior probability density function (PDF) describing the ensemble of all plausible hypocenter locations through Monte Carlo simulation of only the three parameters describing the spatial location of the hypocentre rather than the full inverse problem.

We apply our earthquake location approach to the 1909 M_w 7 Taipei earthquake. Understanding the location of this earthquake is particularly important because it may have been located nearly directly beneath Taipei, a city whose metropolitan area is today home to 7 million people. Our location for this earthquake confirms its proximity to Taipei and suggests that its focal depth was significantly shallower than previous studies have found and, thus, that the seismic hazard to Taipei is proportionally higher.

2 INVERSION DESIGN

We can relate the observed *P*- and *S*-wave arrival times at the *i*th station, $t_p^{(i)}$ and $t_s^{(i)}$, to the earthquake's source location, $\mathbf{x_s}$, by,

$$t_p^{(i)} = t_{ot} + T_p^{(i)}(\mathbf{x_s}) + \epsilon_{clock}^{(i)} + \text{additional errors}, \quad (1)$$

$$t_s^{(i)} = t_{ot} + T_s^{(i)}(\mathbf{x_s}) + \epsilon_{clock}^{(i)} + \text{additional errors}, \quad (2)$$

where $T_p^{(i)}(\mathbf{x_s})$ and $T_s^{(i)}(\mathbf{x_s})$ are the traveltimes between the source location, $\mathbf{x_s}$, and the *i*th station for the *P* and *S* wave, respectively, t_{ot} is the origin time of the earthquake and $\epsilon_{\text{clock}}^{(i)}$ is the clock error at that seismographic station. To calculate the traveltimes, $T_p^{(i)}(\mathbf{x_s})$ and $T_s^{(i)}(\mathbf{x_s})$, we must assume a model of seismic velocities. The origin time, t_{ot}, is a property of the source and is the same for all stations. Note that the clock error, $\epsilon_{\text{clock}}^{(i)}$, is constant for all phase arrivals at the same station. So the more phase arrivals recorded at a station, the more information there is to constrain the clock error separately from the origin time and source location. In principle, the clock error can be determined with just two arrivals. The traveltimes, $T_p^{(i)}$ and $T_s^{(i)}$, are a function of source depth and the source-receiver distance. Thus, for N_{st} stations, our problem has $N_{st} + 4$ unknowns: the latitude, longitude, source depth and origin time of the earthquake; and the clock error at each of the N_{st} stations.

For convenience, let us switch to vector notation. Let $\mathbf{d}$ be a vector of all observed arrival times for all phases at all stations. Thus, if we use *P*- and *S*-wave arrivals,

$$\mathbf{d} = \left(t_p^{(1)}, t_s^{(1)}, ..., t_p^{(N_{st})}, t_s^{(N_{st})}\right)^T, \quad (3)$$

where $(\cdot)^T$ denotes vector transpose.

Then let the predictions for the arrival times for any choice of source location and origin time, $\mathbf{x_s}$ and t_{ot}, be given by,

$$\hat{\mathbf{d}} = \hat{\mathbf{d}}(\mathbf{x_s}, t_{ot}, \boldsymbol{\epsilon}_{\text{clock}}) = t_{ot} + \mathbf{T}(\mathbf{x_s}) + \boldsymbol{\epsilon}_{\text{clock}} \quad (4)$$

so that,

$$\mathbf{d} = \hat{\mathbf{d}} + \boldsymbol{\epsilon}, \quad (5)$$

where $\boldsymbol{\epsilon}$ contains all additional sources of error.

We will solve the inversion for source location given a set of arrival times by using Bayes' theorem (Bayes 1763; O'Hagan 1994),

$$\mathrm{p}(\boldsymbol{\theta}|\mathbf{d}) \propto \mathrm{p}(\mathbf{d}|\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}), \quad (6)$$

where $\boldsymbol{\theta}$ is a vector of model parameters (which for us will be the source location, see below), $\mathbf{d}$ is a vector of data (which for us will be a vector of all arrival times for all phases at all stations), $\mathrm{p}(\cdot)$ denotes a probability density function and $\mathrm{p}(a|b)$ denotes the conditional probability of a given b. $\mathrm{p}(\boldsymbol{\theta}|\mathbf{d})$ is what is usually thought of as the solution to an inverse problem: it is the PDF describing the relative plausibility of all possible values for all model parameters, $\boldsymbol{\theta}$, given a set of observed data, $\mathbf{d}$. $\mathrm{p}(\boldsymbol{\theta})$ is called the prior PDF and is used to constrain the inversion. The prior PDF represents the relative plausibility of different values for our model parameters *a priori*, that is before the introduction of our observations, $\mathbf{d}$. The form of the prior PDF is generally chosen based on *a priori* knowledge about the physical processes being modelled. $\mathrm{p}(\mathbf{d}|\boldsymbol{\theta})$ is the data likelihood function. It describes the misfit between the predictions of a proposed model and our observed data. [For a more complete introduction to Bayesian inference in geophysical inverse problems (see, e.g. Minson *et al.* 2013).] The data likelihood function is often chosen to be a Gaussian, which would be equivalent to an optimization problem where data misfits are assessed using the L_2 norm (e.g. the least-squares method). [In Bayesian analysis, the use of a Gaussian data likelihood is justified by the Principle of maximum entropy, see, e.g. Jaynes (2003) and Beck (2010).]

Let $\mathbf{d}$ be an N_d length vector of observed traveltimes and $\hat{\mathbf{d}}$ be a vector of predicted traveltimes given by eq. (4). Then a Gaussian likelihood function can be written as,

$$\begin{aligned}\mathrm{p}(\mathbf{d}|\boldsymbol{\theta}) &= \mathcal{N}(\mathbf{d}|\hat{\mathbf{d}}, \mathbf{C}_\chi) \\ &= \frac{1}{(2\pi)^{\frac{N_d}{2}}|\mathbf{C}_\chi|^{\frac{1}{2}}} e^{-\frac{1}{2}[\mathbf{d}-\hat{\mathbf{d}}]^T \mathbf{C}_\chi{}^{-1}[\mathbf{d}-\hat{\mathbf{d}}]}, \end{aligned} \quad (7)$$

where $\mathcal{N}(\mathbf{d}|\hat{\mathbf{d}}, \mathbf{C}_\chi)$ denotes the multivariate normal distribution with mean $\hat{\mathbf{d}}$ and covariance $\mathbf{C}_\chi$. $\mathbf{C}_\chi$ represents the errors between our observations and our predictions due to observational noise and errors in our forward model such as errors in the velocity structure used to calculate the predicted traveltimes, but does not include the clock errors that are solved for as part of the inversion.

However, in locating earthquakes from before 1964 (i.e. before the establishment of the World-Wide Standardized Seismographic Stations), we expect to have large outliers in our data, particularly due to some phases being misidentified. (Note that Myers *et al.* 2009, suggested a method for including the possibility that a phase had been misidentified in their methodology for locating modern earthquakes.) An alternative PDF that has longer tails than the normal distribution and thus allows for more outliers is the Laplacian distribution, and has been successfully used for hypocentre location (e.g. Martinsson 2013). This would be equivalent to an L_1 norm for optimization problems (see, e.g. Lomax *et al.* 2009, for discussion of the use of the L_1 norm in earthquake location problems). We can write the Laplacian data likelihood for a set of N_d independent observations (i.e. for diagonal $\mathbf{C}_\chi$),

$$\begin{aligned}\mathrm{p}(\mathbf{d}|\boldsymbol{\theta}) &= \mathcal{L}(\mathbf{d}|\hat{\mathbf{d}}, \mathbf{C}_\chi) \\ &= \prod_{i=1}^{N_d} \frac{1}{\sqrt{2}\sigma_i} e^{-\frac{\sqrt{2}}{\sigma_i}|\mathbf{d}_i - \hat{\mathbf{d}}_i|}, \end{aligned} \quad (8)$$

where σ_i^2 is the ith diagonal element of $\mathbf{C}_\chi$. Thus, the Laplacian data likelihood has the same variance as the Gaussian data likelihood (eq. 7) for any assumed covariance matrix of errors, $\mathbf{C}_\chi$.

Gaussian and Laplacian PDFs are just two potential choices for the data likelihood. Foremost among the many other possibilities is the data likelihood proposed by Jeffreys to minimize erroneous picks of phase times for earthquake location (Jeffreys 1932; Buland 1986). However, we believe that the main source of error in historical earthquake location comes from clock errors, and our goal is to explore how to best minimize the effects of those errors.

Formally, we have $N_{st} + 4$ unknowns (latitude, longitude, depth, origin time and a clock error at each of N_{st} stations), which could be a large number of unknowns if the number of stations is large. However, we will employ a shortcut to make the problem appear to have only the three unknowns defining the source location, $\mathbf{x_s}$: latitude, longitude and depth. Rather than determine the posterior PDF describing all plausible values for all $N_{st} + 4$ unknowns, we will only explore the space of all plausible source locations, $\mathbf{x_s}$, and, for each potential source location, we will then remove the optimal origin time shift and clock error shifts (determined by a least-squares fit) before evaluating the posterior probability. Specifically, for each candidate source location, $\mathbf{x_s}$, we evaluate,

$$\mathbf{r} = \mathbf{d} - \mathbf{T}(\mathbf{x_s}) \tag{9}$$

this is the vector of residuals between the observed traveltimes and predicted traveltimes without clock errors or an origin time shift. We then solve the linear equation,

$$\mathbf{r} = \mathbf{G}\mathbf{m}, \tag{10}$$

where $\mathbf{m}$ is an $N_{st} + 1$ length vector containing the clock error for each station as well as the origin time. Thus, if $\mathbf{d}$ is arranged as in eq. (3), $\mathbf{G}$ has the form,

$$\mathbf{G} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix}. \tag{11}$$

We identify the first column of $\mathbf{G}$ as the origin time shift and each succeeding column as the clock error for one station. The rows appear in pairs in this example because we have both *P*- and *S*-wave arrival information and the clock error will be identical for both phase picks.

Substituting the least-squares solution to eq. (10) into eq. (4), we find,

$$\hat{\mathbf{d}} = \mathbf{T}(\mathbf{x_s}) + \mathbf{G}(\mathbf{G}^T {\mathbf{C}_\chi}^{-1} \mathbf{G})^{-1} \mathbf{G}^T {\mathbf{C}_\chi}^{-1} \mathbf{r}. \tag{12}$$

Given that we have extracted the origin time and clock errors from our inversion by replacing them with a least-squares fit to the traveltime residuals, the only unknowns we have left are those describing the earthquake location, $\mathbf{x_s}$. Thus, $\theta = \mathbf{x_s}$ in eq. (6). For our prior distribution $p(\theta)$, we use broad Gaussians centred on an initial guess for the hypocentre location. However, since depth must be a positive quantity, we enforce positivity by parameterizing the source location as θ = (latitude, longitude, ln depth), and thus our prior PDF on source depth is a Gaussian on ln depth or a log-normal distribution on depth.

There are several potential advantages to using a log-normal distribution for the prior PDF on source depth. Besides it satisfying the need to eliminate negative hypocentral depths, its exponential decay with increasing depths makes the plausibility of a very deep earthquake vanishingly small without the need to assign a specific lower limit to potential source depths. Further, it allows us to introduce information about which range of source depths is most plausible given our *a priori* information about the depth of the earthquake under study from independent studies, known fault structures, or observed seismicity. Alternatively, we could have, for example, adopted a uniform prior PDF ranging from the Earth's surface to some lower seismogenic limit, or we could have constructed an empirical PDF based on the depths of all observed earthquakes in the region. Whatever our choice for prior PDF, the best way to evaluate its effectiveness is to examine the posterior PDFs from inverting synthetic data using that prior PDF. All of our tests in Section 3 were able to successfully recover the input hypocentral depth using a log-normal prior PDF, and thus it appears that the log-normal distribution is an effective choice for the prior PDF.

We then simulate the posterior PDF describing the relatively plausibility of all possible source locations by using Markov Chain Monte Carlo methods to draw random samples distributed according to the posterior PDF. Our sampling algorithm is the cascading adaptive transitional metropolis in parallel algorithm (Minson *et al.* 2013), an efficient, parallelized, self-optimizing sampler which combines elements of simulated annealing and genetic algorithms with many parallel instances of the Metropolis algorithm (Metropolis *et al.* 1953; Chib & Greenberg 1995). The Metropolis algorithm, in turn, uses a random walk (or Markov chain) to explore the space of possible values for each model parameter. For all of the following examples, the solutions were computed with 1000 parallel Markov chains each of which was 50 random walk steps long. For each candidate sample in each Markov chain, $\mathbf{T}(\mathbf{x_s})$ was evaluated by looking up the appropriate traveltime in an AK135 traveltime table (Kennett *et al.* 1995). Due to the extremely small number of free parameters being searched (three) and the speed of the forward model (which is a table lookup), the entire inverse problem takes just seconds to solve on a personal computer.

3 APPLICATION TO THE 1909 TAIPEI EARTHQUAKE

The magnitude 7 Taipei earthquake of 1909 April 14, was studied extensively in Kanamori *et al.* (2012) because it is an important event in assessing seismic hazard in the Taipei metropolitan area (Fig. 1). In that paper, the authors relocated the earthquake hypocentre using the JLoc direct search software package (Lee & Dodge 2007), and compared their results to the hypocentral estimate published in Gutenberg & Richter (1954) and an epicentre reported by the local Taihoku observatory (Taihoku Meteorological Observatory 1936; Table 1). For our prior distribution, $p(\theta)$, we use Gaussian distributions centred roughly around the Kanamori *et al.* (2012) hypocentre: latitude 25°N ± 5°, longitude 121.5°E ± 5°, and a log-normal distribution on depth such that the mean of the corresponding normal distribution is 50 km (Table 2).

Kanamori *et al.* (2012) made an extensive effort to collect a worldwide data set of phase picks and seismograms for the 1909 Taipei earthquake. They utilized arrival times from a total of 41 stations (including eight local stations in Taiwan) to locate the earthquake's hypocentre (see table B2 in Kanamori *et al.* 2012). In the present paper, we used only 15 stations that were considered to have reliable

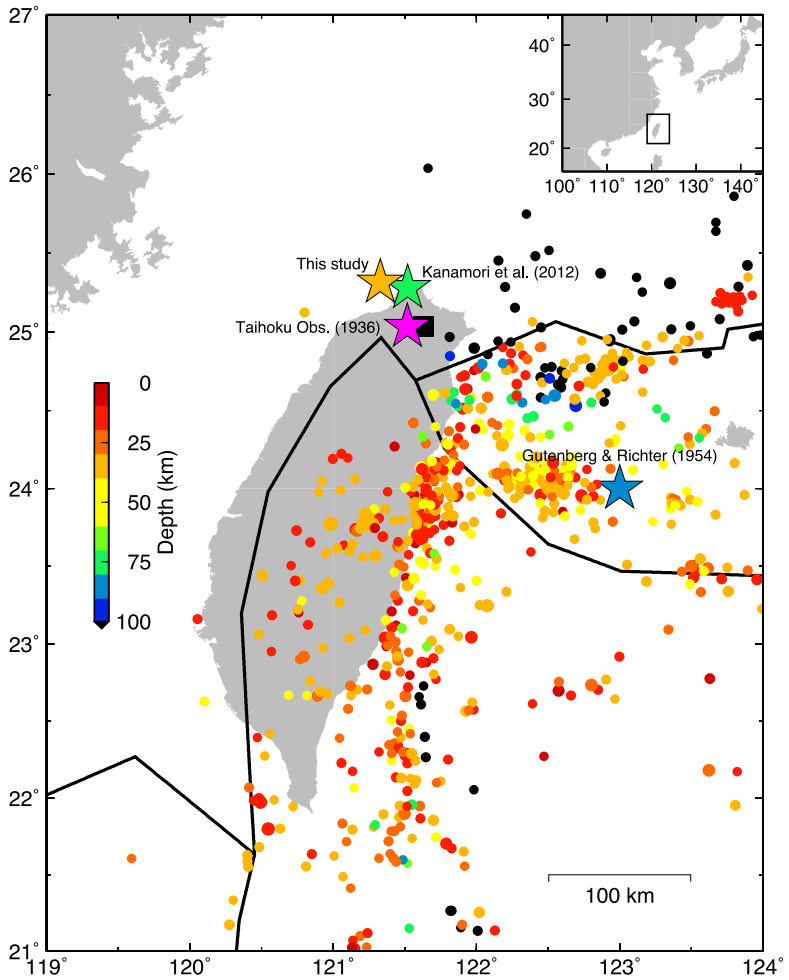

Figure 1. Comparison of locations for the 1909 Taipei earthquake from this study and others to the background seismicity. All earthquake symbols except for the Taihoku observatory's epicentre (which lacks a source depth) are coloured by hypocentral depth. Plate boundaries are plotted with black lines. Background seismicity is all M5 and larger earthquakes in the U.S. Geological Survey's Comprehensive Catalogue (ComCat) 1973–present. Black square indicates the location of Taipei.

Table 1. Published locations for the 1909 Taipei earthquake.

Source	Latitude	Longitude	Depth
Kanamori *et al.* (2012)	25.28°N	121.52°E	75 km
Gutenberg & Richter (1954)	24°N	123°E	80 km
Taihoku Meteorological Observatory (1936)	25.033°E	121.517°E	–

Table 2. Prior distributions.

Parameter	Prior PDF
Latitude	$\mathcal{N}(25°\text{N}, 5°)$
Longitude	$\mathcal{N}(121.5°\text{E}, 5°)$
ln Depth	$\mathcal{N}$[(3.912 ln(km), 1 ln(km)]

readings based on the fact that their seismograms could be read at a precision of about 1 s, yielding approximate 3 s residual uncertainties according to Kanamori *et al.* (2012). Thus, the data set for the inversions presented in this paper consists of the *P*- and *S*-wave arrival times at 15 stations (GRA, GTT, JEN, LEI, POT, tHEN, TIF, tKLG, tPNG, tTAP, tTCU, tTNN, tTTN, UPP, ZKW) using the phase picks of Kanamori *et al.* (2012) with assumed uncertainties of $\sigma = 3$ s.

We explored three synthetic test cases before inverting the real observed traveltimes for the 1909 Taipei earthquake. In all of the synthetic cases, we use the Kanamori *et al.* (2012) hypocentre as the true source location. In the first synthetic case, we simply invert perfect noise-free synthetic *P*- and *S*-wave traveltimes. In the second case, we apply a clock error to the synthetic observations at each station. In the third case, we add to these same clock errors an extra noise source which represents all other possible sources of error such as mistakes in phase picking and inaccuracies in the traveltime velocity model. This last synthetic data set was calculated by adding to each phase pick both a clock error and a draw from the uniform distribution $\mathcal{U}(-\frac{1}{4}\sigma, +\frac{3}{4}\sigma)$ where σ is the assumed observational error of 3 s. We chose a non-zero-mean uniform distribution because most phase pick errors are biased towards delays due to the emergent nature of some seismic phases.

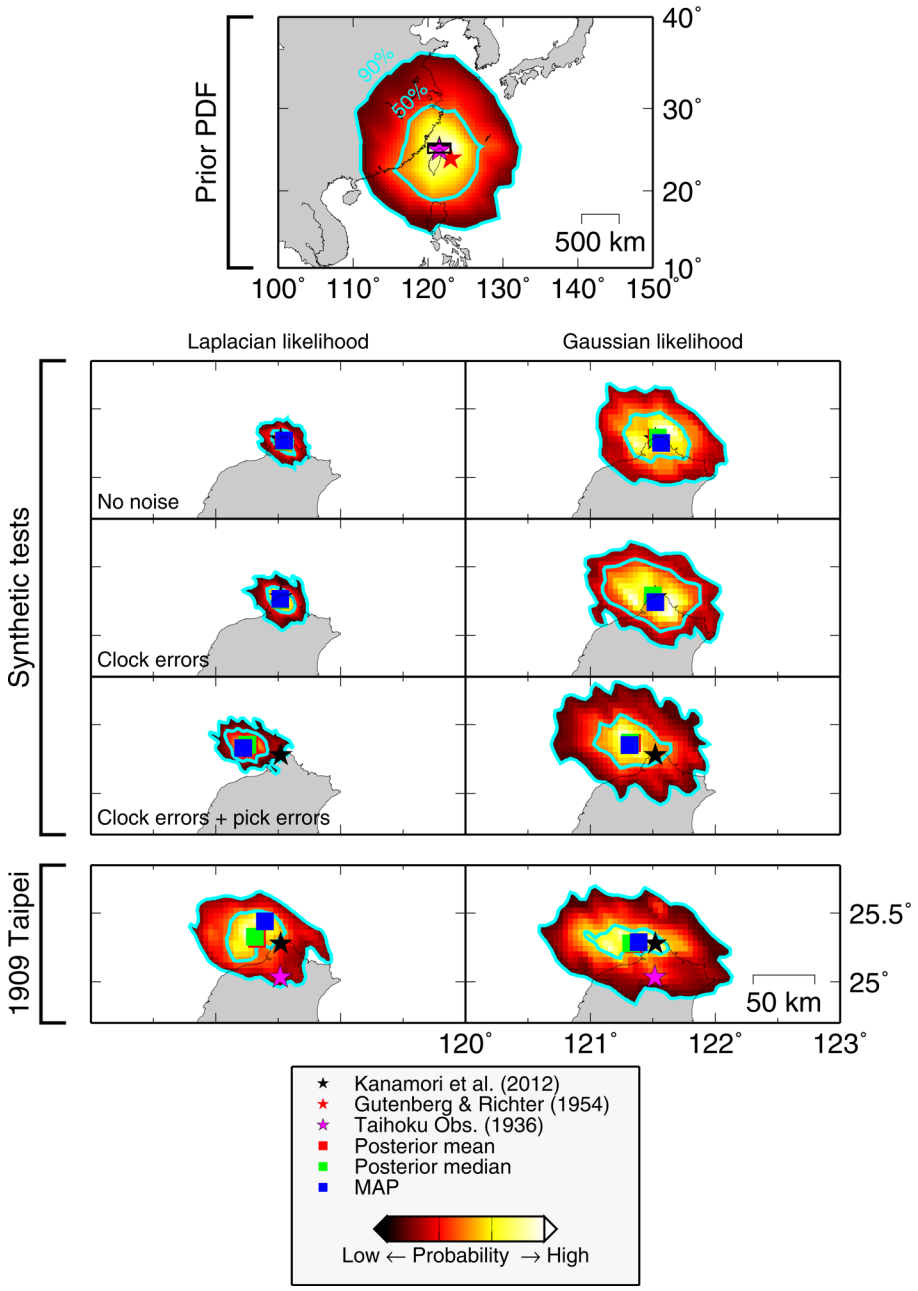

Figure 2. Posterior PDF of latitude and longitude for all synthetic tests as well as inversion using observed traveltimes for the 1909 Taipei earthquake. The solutions in the left column were computed using a Laplacian likelihood function. The right column employs a Gaussian likelihood function. The posterior mean, posterior median and maximum *a posteriori* (MAP) epicentres are marked with coloured squares. Published locations are shown with stars in the top and bottom rows. Contours of samples of the prior PDF are shown for reference in the top row. Colour indicates relative probability density. Cyan contours show 50 and 90 per cent credibility regions. The input hypocentre for the synthetic tests is the Kanamori *et al.* (2012) location. The small box shown in the plot in the top row marks the spatial extent of the plots in the lower rows. The city of Taipei has the same location as the Taihoku observatory epicentre (pink star).

The posterior distributions for the latitude and longitude of source location are shown in Fig. 2. The top row shows samples of the prior distribution, which cover a much larger geographic range than the inversion results (posterior PDFs) from any of the synthetic tests or the inversion of real data. (Note that the map scale of the prior PDF subplot is much larger than the other rows of Fig. 2.) This difference in ranges shows that, using only the *P*- and *S*-wave arrival times at a limited number of early seismographic stations, it is possible to resolve the source location of the 1909 Taipei earthquake.

In all of the synthetic tests, despite the small quantity and low quality of data, the inversion returned an ensemble of locations that converge towards the input source epicentre (shown with a black star). We can look at statistics of the posterior PDF, such as the mean of the posterior samples, the median of the posterior samples and the single model with the maximum *a posteriori* probability (MAP). The epicentre location derived from these three statistics are shown for every inversion in Fig. 2. In the synthetic test with noise-free data, the posterior mean, posterior median and MAP are nearly perfect matches for the input epicentre location. As sequentially more noise sources are added to the synthetic data, the posterior mean, posterior median and MAP become increasingly biased from the true epicentre, but these errors are small and the input epicentre

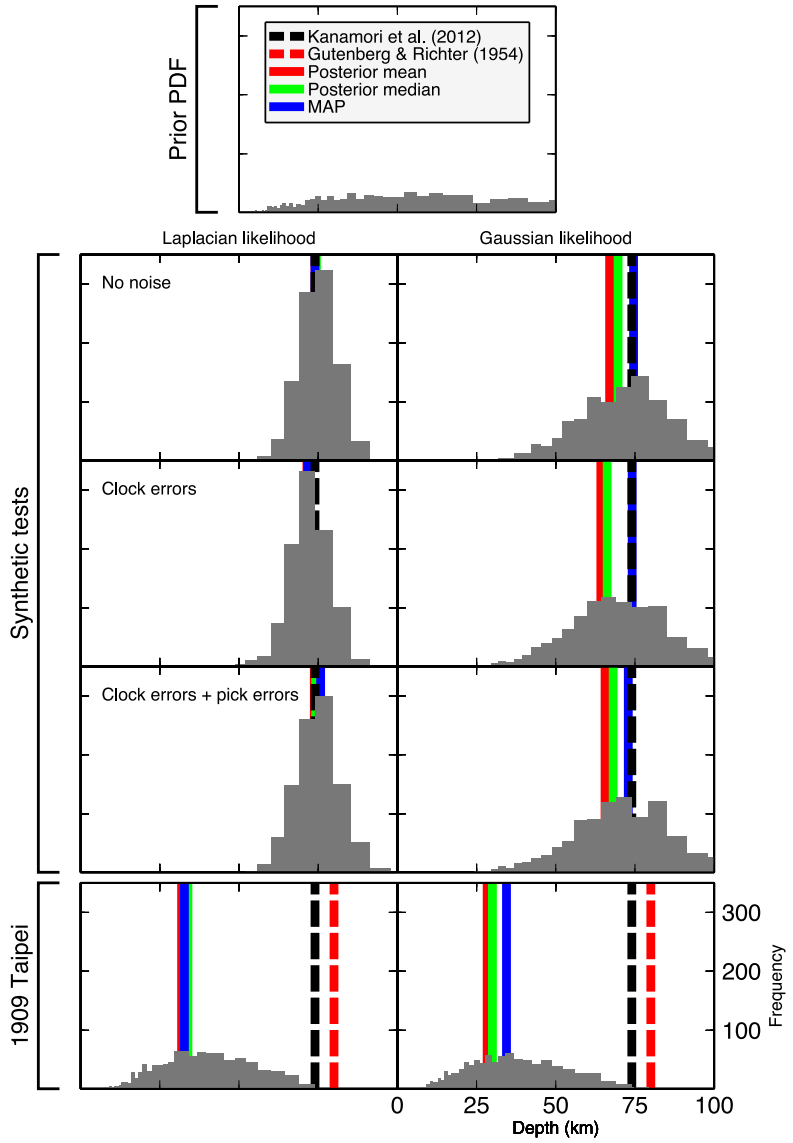

Figure 3. Posterior histograms of depth for all inversions. The solutions in the left column were computed using a Laplacian likelihood function. The right column employs a Gaussian likelihood function. Published focal depths are shown for comparison. The top row of plots show the distribution of samples of the prior PDF before inversion. The next three rows of plots show the posterior PDFs after inversion for synthetic tests using perfect synthetic data, synthetic data containing clock errors added to perfect phase picks and synthetic data containing both clock errors and errors on the phase picks, respectively. The input hypocentre for the synthetic tests is the Kanamori *et al.* (2012) location. The bottom row of plots show the results of our inversion for the 1909 Taipei earthquake. Note that our inversions for the location of the 1909 Taipei earthquake robustly prefer a shallower hypocentral depth than previous studies. Since the synthetic tests are capable of recovering the deeper hypocentre of Kanamori *et al.* (2012), the shallow hypocentre from this study does not appear to be an artifact of the inversion methodology.

is well within the spatial extent of the posterior PDF. Thus, in all test cases, our inversion approach produces a posterior PDF which recovers the input epicentre location within the uncertainties on the inferred model parameters where the posterior uncertainties are shown by the broadness of the posterior PDF.

Both likelihood functions produce acceptable results and, in fact, produce very similar solutions for the latitude and longitude of the earthquake. The only potentially significant difference in the performance of the two likelihood functions is when we consider the ensemble of possible hypocentral depths (Fig. 3). Here, the Laplacian likelihood function recovers the input depth almost perfectly for the synthetic tests. The Gaussian likelihood is somewhat less capable of recovering the input source depth in the synthetic tests, and the resulting posterior PDF is much broader, making it harder to identify an optimal source depth. However, these results are likely controlled by the type of noise used in creating the synthetic data set. We used a uniform distribution to simulate errors on the phase picks. This is a heavy-tailed distribution and thus may favour inversions using a Laplacian data likelihood instead of a Gaussian likelihood. The differences between the Laplacian and Gaussian results diminish significantly when applied to real data, and a posterior analysis of the residuals show that the distribution of errors are equally well described by a Gaussian or a Laplacian PDF according to the test statistic of Kundu (2005). Thus, either of these two data likelihoods appear to be equally good choices for describing the observed data and errors.

Potential trade-offs between epicentre location and focal depth can be assessed from the cross-sections of the posterior PDF

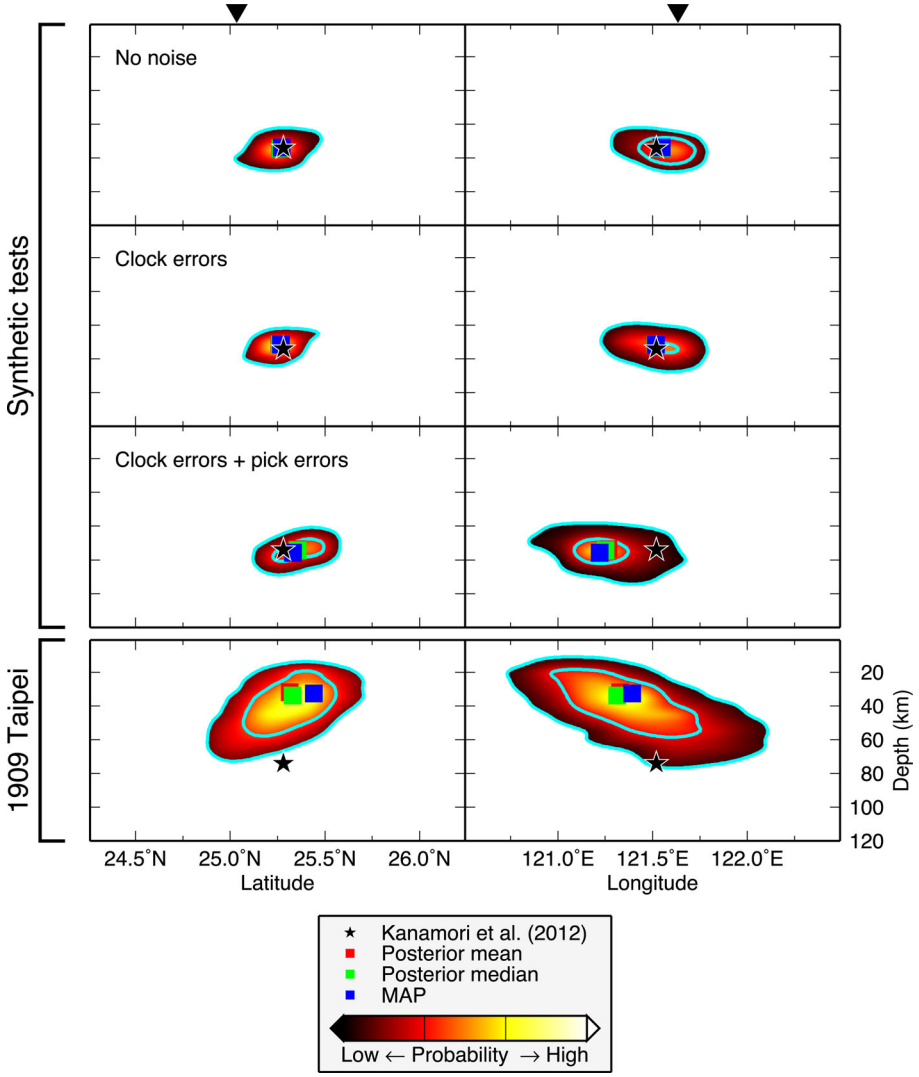

Figure 4. Cross-sections of posterior PDFs for inversions with Laplacian likelihood function. The posterior PDF of hypocentral depth is plotted versus hypocentral latitude (left) and hypocentral longitude (right). Colour indicates relative probability density. Cyan contours show 50 and 90 per cent credibility regions. The input hypocentre for the synthetic tests is the Kanamori *et al.* (2012) location, shown with a star. The mean, median, and MAP hypocentres are shown with squares. The location of the city of Taipei is marked by a triangle.

presented in Figs 4 and 5. There appears to be a slight trade-off in the hypocentre location of the 1909 Taipei earthquake with shallower source depths preferred for more northwestern epicentres and deeper focal depths for southeastern hypocentres. Although this trade-off is slight and thus hard to discern for the synthetic tests which resulted in tight posterior distributions, it appears that this trade-off may exist in the results of the synthetic tests as well, suggesting that it is an effect of the station distribution and not the earthquake source or some difference between the noise added to our synthetic phases picks and the observed noise for the 1909 Taipei earthquake.

For comparison, we ran the test case of perfect data and the case of synthetic data with clock errors through the JLoc inversion program. For the perfect noise-free synthetic data, JLoc did an excellent job of recovering the hypocentre, and it was able to recover the epicentre for the case without clock errors. However, once clock errors were introduced into the synthetic data, the JLoc solution put the depth of the earthquake at 0 km.

The inversion results for the origin time shift relative to the assumed origin time of 1909/04/14 19:53:52.5 are shown in Fig. 6. The synthetic examples use an origin time shift of zero, and the inversion does a good job of recovering zero-shift for the noise-free case and the example with clock errors only. When additional noise with a systematic bias is added to the synthetic observations, the inversion results are not quite as good, but the mean posterior origin time shifts are still less than two seconds. For real data, the inversion suggests that the supposed origin time for the 1909 Taipei earthquake may be wrong by about one second, but the uncertainties on these results are such that it is probably not a robust finding.

We verify that our inversion methodology can correctly recover the clock error at each station by comparing the posterior mean estimate for the clock errors at each station in each synthetic test to the input errors used to generate the synthetic data (Fig. 7). For the case with only clock errors, both likelihood functions do an excellent job of estimating the clock error at each station. For the synthetic test which includes additional noise, the estimates are

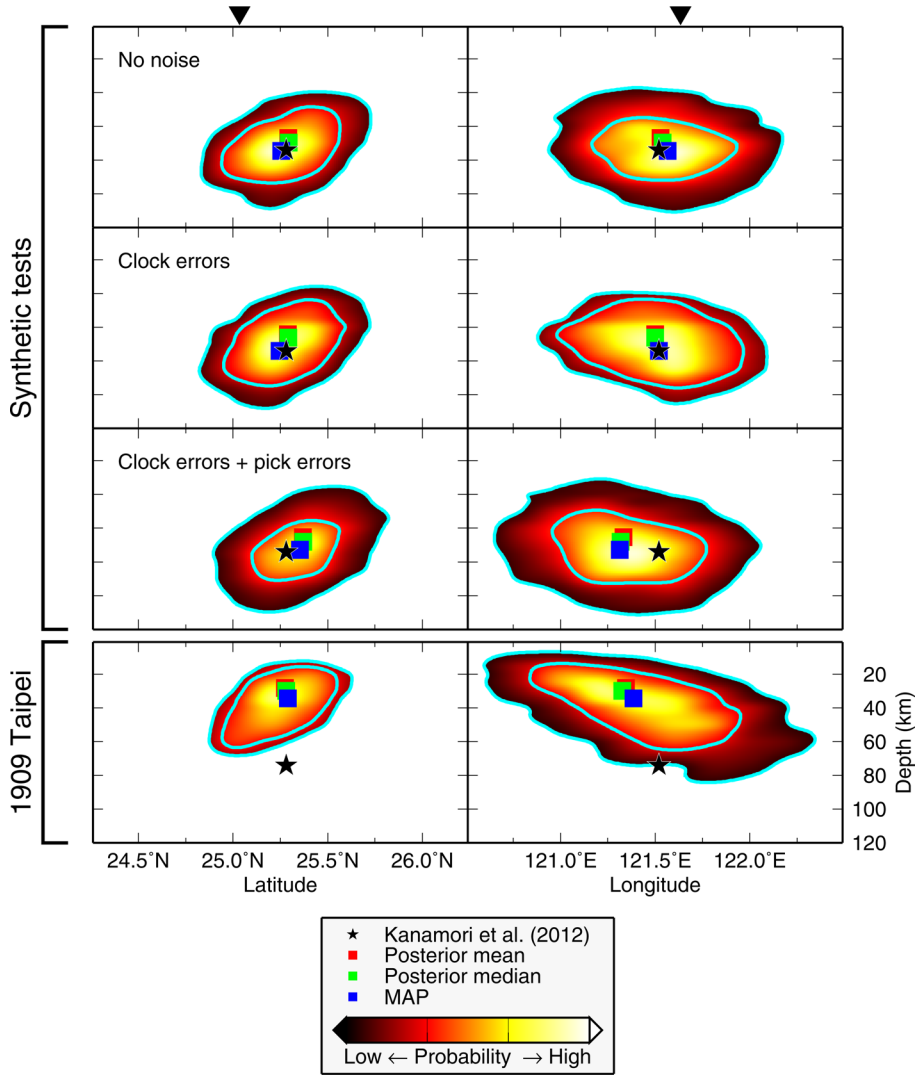

Figure 5. Cross-sections of posterior PDFs for inversions with Gaussian likelihood function. Same as Fig. 4 for posterior PDFs calculated using a Gaussian data likelihood.

not quite as good, but still quite accurate. This indicates that our inversion for clock errors is reliable.

4 DISCUSSION

If we compare our epicentre location and hypocentral depth for the 1909 Taipei earthquake (bottom rows of Figs 2 and 3, respectively) to the previously published locations, we see that our epicentre location is very close to the Kanamori *et al.* (2012) which, in turn, is very close to the Taihoku Meteorological Observatory (1936) epicentre. Both the Kanamori *et al.* (2012) and Taihoku Meteorological Observatory (1936) epicentres are within our posterior PDF, although the Kanamori *et al.* (2012) location is more consistent with our results. (Compare Table 1 to Tables 3 and 4.) In contrast, the Gutenberg & Richter (1954) epicentre is outside of our posterior PDF and thus our inversion excludes that epicentre location. However, Gutenberg & Richter (1954) assigned their location for the 1909 Taipei earthquake nominal errors of three degrees in latitude and longitude and Kanamori *et al.* (2012) note that the Taihoku Meteorological Observatory (1936) epicentre is precisely the location of the city of Taipei and thus the Taihoku observatory may have simply assigned the 1909 earthquake location to the latitude and longitude of the nearest city. Given these uncertainties on the Gutenberg & Richter (1954) and Taihoku Meteorological Observatory (1936) locations, we may consider their epicentres consistent with both the Kanamori *et al.* (2012) location and the results of this study.

Our hypocentral depth (Fig. 3) is significantly shallower than either of the previously reported hypocentres (Gutenberg & Richter 1954; Kanamori *et al.* 2012). Based on synthetic tests (which used the source depth of Kanamori *et al.* (2012) to generate the synthetic phase picks), we know that our inversion is capable of recovering deeper hypocentres. Thus, we do not think that our shallower hypocentre is an artifact of our inversion methodology. Rather, we think that the 1909 Taipei earthquake was significantly shallower than previously believed which could mean that the seismic hazard in Taipei is higher than thought.

Taipei experienced significant damage from the 1909 Taipei earthquake: more than a thousand homes were damaged and 60 people were injured or killed (Taihoku Meteorological Observatory 1936; Kanamori *et al.* 2012). Kanamori *et al.* (2012) suggested that the 1909 earthquake must have been enriched in high-frequency energy to explain how a relatively deep event could cause such strong ground shaking near Taipei. However, a simpler explanation is that,

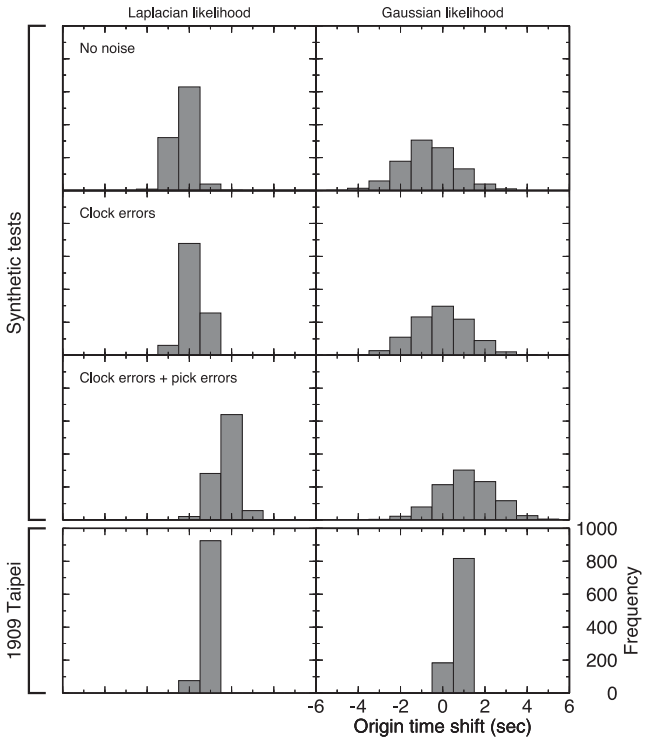

Figure 6. Posterior histograms of origin time shift for all inversions. The solutions in the left column were computed using a Laplacian likelihood function. The right column employs a Gaussian likelihood function. The top three rows of plots show the posterior PDFs for synthetic tests using perfect synthetic data, synthetic data containing clock errors added to perfect phase picks, and synthetic data containing both clock errors and errors on the phase picks, respectively. The synthetic data sets were made with a time shift of zero seconds and, if the inversion were perfect, the posterior histograms would be tightly peaked around zero. The origin time shift for the 1909 Taipei earthquake is calculated relative to the origin time of Kanamori *et al.* (2012).

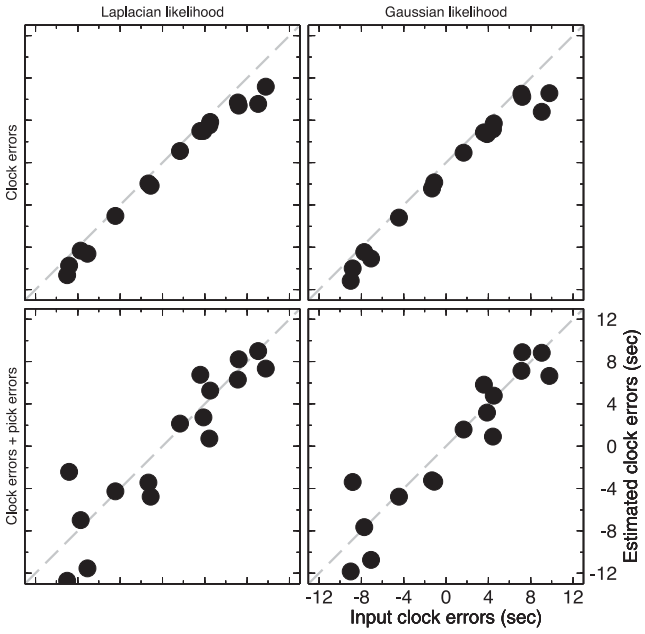

Figure 7. Comparison of estimated clock errors to the input clock errors used to generate synthetic data. The estimated clock errors are the means of the posterior samples. The posterior mean clock error at each station is plotted against the input clock error used to construct the synthetic phase picks at that station. (Thus, each circle represents the clock error at a different station and, if the clock errors were perfectly recovered, the circles would lie on the dashed line.) The solutions in the left column were computed using a Laplacian likelihood function. The right column employs a Gaussian likelihood function. The top row of plots is for inversions using synthetic data containing clock errors added to perfect phase picks. The bottom row is for inversions with synthetic data containing both clock errors and errors on the phases picks.

as the results of this study suggest, the earthquake had a shallow focus.

The observed seismicity in the region of our hypocentre, the Kanamori *et al.* (2012) hypocentre, and the Taihoku Meteorological Observatory (1936) epicentre is sparse, making it difficult to use the observed seismicity to assess which of these three source locations is more probable. Kanamori *et al.* (2012) preferred a focal depth of 75 km, although they note that the 1909 Taipei record recorded at GTT was comparable to a record from the 2001 Geiyo earthquake whose source depth was just 47.4 km. While 75 km is deeper than the range of source depths allowed by the 95 per cent credibility interval on our hypocentre location, the focal depth of the Geiyo earthquake does lie in our 95 per cent credibility bounds (Fig. 3; Tables 3 and 4). (In Bayesian analysis, the 95 per cent credibility interval for a parameter describes a region such that there is a 0.95 probability that the value of that parameter lies in that region.) However, our preferred source depth is closer to 30 km, which is significantly shallower than the apparent depth of the subducted Philippine Sea Plate in the source region of the 1909 Taipei earthquake as seen by seismic tomography (Wu *et al.* 2009). This suggests that the 1909 Taipei earthquake might have been a deep crustal event rather than an intraslab earthquake as concluded by Kanamori *et al.* (2012), but more work is needed to confirm such a hypothesis.

In addition to modelling the hypocentre location of the 1909 Taipei earthquake, we also determined the origin time of the earthquake and the clock error at each seismographic station. The posterior distribution of estimated clock errors for the 1909 Taipei earthquake is shown in Figs 8 and 9. We note that the magnitude of the estimated errors roughly mirrors the reputation that each station has for quality. Established seismographic stations like GTT and UPP have small clock errors, while the local Taiwan stations can have quite large errors. Thus, our inferred clock errors make intuitive sense. Among the local Taiwan stations, the main Taihoku observatory located in Taipei (tTAP) might be thought to have accurate timekeeping. However, our results suggest that while the Taihoku observatory had one of the smallest clock errors in Taiwan, its clock error was still almost 10 s. In fact, this result is not surprising since the Taihoku observatory is a meteorological, not astronomical, observatory and thus might not have calibrated their clock accurately.

Our inferred origin time for the 1909 Taipei earthquake is only about a second later than the Kanamori *et al.* (2012) origin time. However, we do not believe that this potential origin time shift is a robust feature of the source location inversion. Further, because we model the origin time as the average time shift at all stations (eq. 11) and we expect the errors on phase picks to be biased towards delays because of the emergent nature of some seismic phases, there is the potential to map systematic bias in the observed phase picks into a delayed origin time. However, the fact that our inversion did not produce much delay in the inferred origin time relative to the Kanamori *et al.* (2012) origin time may be an indicator that our methodology is approximately as capable at estimating the correct origin time as the JLoc approach used by Kanamori *et al.* (2012).

Table 3. Inversion results for 1909 Taipei earthquake using Laplacian data likelihood.

	Latitude	Longitude	Depth
Posterior mean	25.315°N	121.329°E	31.9 km
Posterior median	25.329°N	121.313°E	33.8 km
Maximum *a posteriori*	25.442°N	121.393°E	32.8 km
95 per cent credibility interval	(24.974°N, 25.596°N)	(120.812°E, 121.919°E)	(11.0 km, 65.4 km)

Table 4. Inversion results for 1909 Taipei earthquake using Gaussian data likelihood.

	Latitude	Longitude	Depth
Posterior mean	25.274°N	121.346°E	28.3 km
Posterior median	25.280°N	121.327°E	30.0 km
Maximum *a posteriori*	25.289°N	121.386°E	34.5 km
95 per cent credibility interval	(24.924°N, 25.576°N)	(120.739°E, 122.043°E)	(8.0 km, 65.5 km)

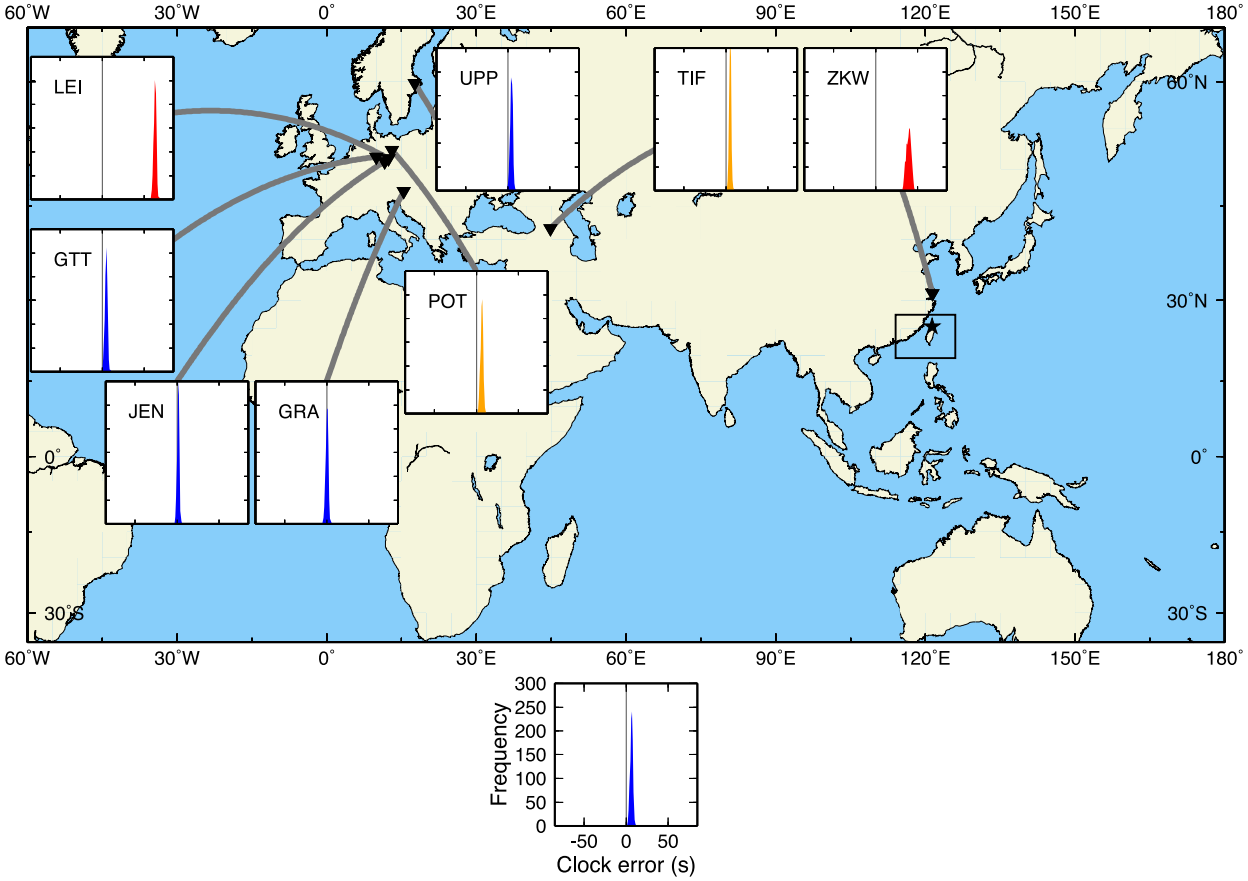

Figure 8. Estimated clock errors at global stations from inversion for the 1909 Taipei earthquake using Laplacian likelihood function. Histograms of the estimated clock errors associated with all hypocentres in the posterior PDF and are coloured according to the posterior mean of the distribution of errors at each station. (These histograms can be viewed as unnormalized posterior PDFs of clock errors.) Posterior mean errors less than 5 s are coloured blue, mean errors between 5 and 10 s are orange, and mean errors in excess of 10 s are red. Estimated clock errors for stations in region marked by black box are shown in Fig. 9.

5 CONCLUSIONS

We have demonstrated an approach that uses both *P*- and *S*-wave arrival times to simultaneously solve for the source location and origin time of historical earthquakes as well as clock errors at each station. The ability to simultaneously solve for the hypocentre and potential clock errors could be of particular use for studying older events from the late 19th and early 20th centuries when clock errors were a more common problem and the scarcity of available data makes it difficult to throw out data that are suspected of containing a clock error.

This is a preliminary proof-of-concept study for which we used *P*- and *S*-wave arrivals only. However, in the future, we could consider using other seismic phases as well. The addition of more phases at the same station can only help to constrain the source location, origin time and clock errors. Further, we have only used a naive Gaussian prior distribution on our source location. In future studies, we could use knowledge about plate boundaries, fault geometries and catalogue earthquake locations to provide a more informative prior which can improve our earthquake locations.

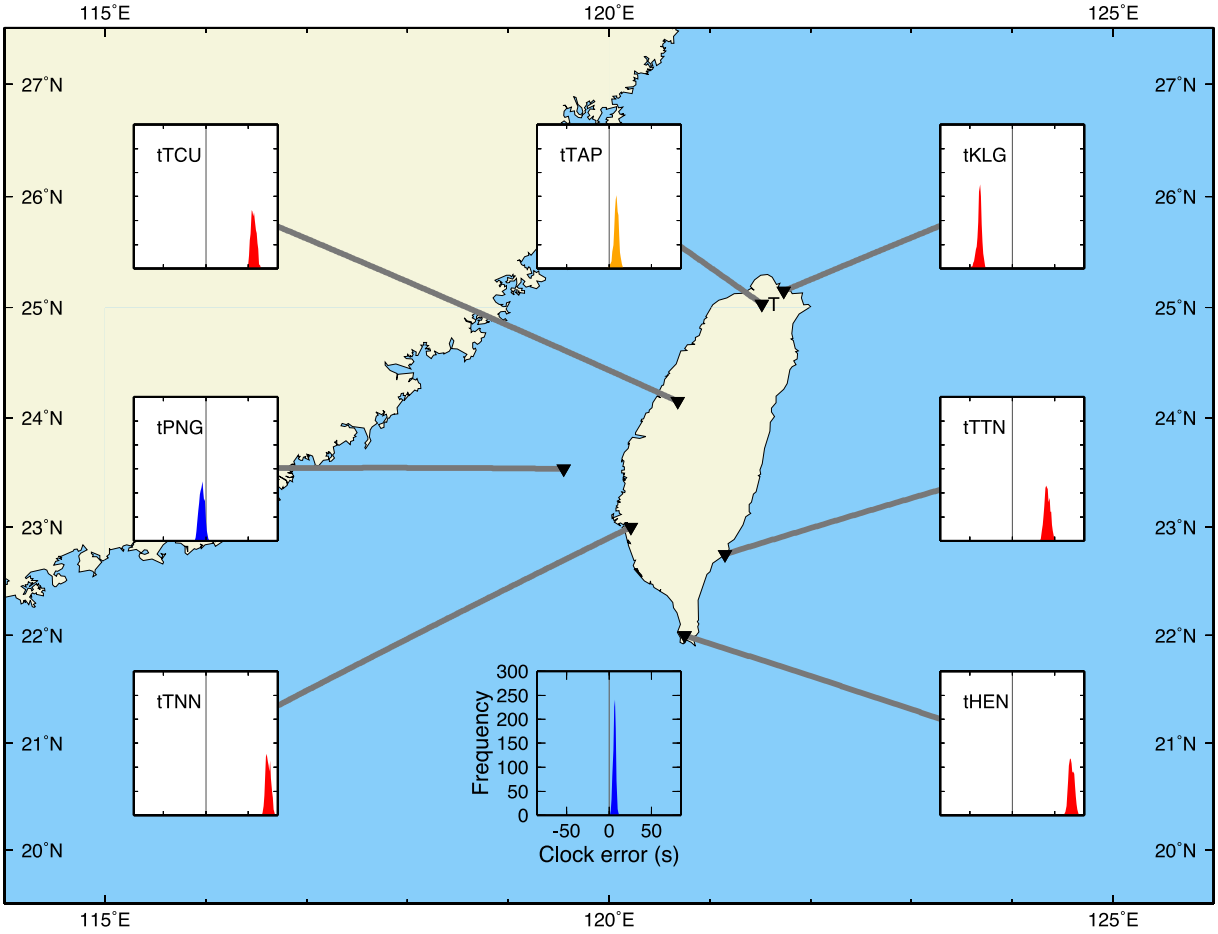

Figure 9. Estimated clock errors at local stations from inversion for the 1909 Taipei earthquake using Laplacian likelihood function. All symbols and colours are the same as Fig. 8. The location of the city of Taipei is shown with a T.

Even using the simplistic approach presented here, we were able to locate the 1909 Taipei earthquake at least as well as existing studies and to solve for the origin time and clock errors in addition. The clock errors we found are consistent with our intuition about which seismographic stations produced higher quality observations and which did not. Our results for source depth, although accompanied with large uncertainties, seem to indicate that the 1909 Taipei earthquake might have been shallower than previously thought. This result, if true, could significantly increase the shaking hazard faced by the Taipei metropolitan area.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Dr. Fred Klein and Dr. Jim Savage for helpful reviews of our original manuscript draft. We thank Prof. Duncan Agnew, the editor, Dr. Anthony Lomax, the reviewer, and one anonymous reviewer for their careful and detailed reviews of our submitted manuscript. The comments and suggestions we received greatly improved our manuscript.

REFERENCES

Bayes, T., 1763. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S., *Phil. Trans. R. Soc.*, **53**, 370–418.

Beck, J., 2010. Bayesian system identification based on probability logic, *Structu. Control Health Monitor.*, **17**, 825–847.

Bolt, B.A., 1960. The revision of earthquake epicentres, focal depths and origin-times using a high-speed computer, *Geophys. J. Int.*, **3**(4), 433–440.

Buland, R., 1986. Uniform reduction error analysis, *Bull. seism. Soc. Am.*, **76**(1), 217–230.

Chib, S. & Greenberg, E., 1995. Understanding the Metropolis-Hastings algorithm, *Am. Stat.*, **49**(4), 327–335.

Engdahl, E.R. & Gunst, R.H., 1966. Use of a high speed computer for the preliminary determination of earthquake hypocenters, *Bull. seism. Soc. Am.*, **56**(2), 325–336.

Geiger, L., 1910. Herdbestimming bei erdbeben aus den ankunftszeiten, *Koniglichen Geselhchaft der Wissmchaften zu Gottingen*, **4**, 331–349.

Gutenberg, B. & Richter, C.F., 1954. *Seismicity of the Earth and Associated Phenomena*, 2nd edn, Princeton Univ. Press.

Jaynes, E., 2003. *Probability Theory: The Logic of Science*, Cambridge Univ. Press.

Jeffreys, H., 1932. An alternative to the rejection of observations, *Proc. R. Soc. Lond., A*, **137**(831), 78–87.

Jordan, T.H. & Sverdrup, K.A., 1981. Teleseismic location techniques and their application to earthquake clusters in the south-central Pacific, *Bull. seism. Soc. Am.*, **71**(4), 1105–1130.

Kanamori, H., Lee, W. H.K. & Ma, K.-F., 2012. The 1909 Taipei earthquake – implication for seismic hazard in Taipei, *Geophys. J. Int.*, **191**(1), 126–146.

Kennett, B.L.N., Engdahl, E.R. & Buland, R., 1995. Constraints on seismic velocities in the Earth from traveltimes, *Geophys. J. Int.*, **122**(1), 108–124.

Kundu, D., 2005. Discriminating between normal and Laplace distributions, in *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, pp. 65–79, eds Balakrishnan, N., Kannan, N. & Nagaraja, H.N., Springer-Verlag.

Lee, W.H.K. & Dodge, D.A., 2007. *Development of a direct search software package for poorly constrained earthquakes, Tech. rep.*, Seismology Technical Report, Central Weather Bureau, Taiwan. Available at: http://www.iris.edu/seismo/quakes/1909Taiwan/Lee-Dodge_2007.pdf (last accessed 16 February 2014).

Lee, W.H.K. & Stewart, S.W., 1981. *Principles and Applications of Microearthquake Networks*, Advances in Geophysics, Supplement 2, Academic Press.

Lomax, A., 2005. A reanalysis of the hypocentral location and related observations for the great 1906 California earthquake, *Bull. seism. Soc. Am.*, **95**(3), 861–877.

Lomax, A., Michelini, A. & Curtis, A., 2009. Earthquake location, direct, global-search methods, in *Encyclopedia of Complexity and Systems Science*, pp. 2449–2473, ed. Meyers, R.A., Springer-Verlag.

Martinsson, J., 2013. Robust Bayesian hypocentre and uncertainty region estimation: the effect of heavy-tailed distributions and prior information in cases with poor, inconsistent and insufficient arrival times, *Geophys. J. Int.*, **192**(3), 1156–1178.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E., 1953. Equation of state calculations by fast computing machines, *J. Chem. Phys.*, **21**(6), 1087–1092.

Minson, S.E., Simons, M. & Beck, J.L., 2013. Bayesian inversion for finite fault earthquake source models I – theory and algorithm, *Geophys. J. Int.*, **194**(3), 1701–1726.

Mosegaard, K. & Tarantola, A., 2002. Probabilistic approach to inverse problems, in *International Handbook of Earthquake and Engineering Seismology, Part A*, pp. 237–265, eds Lee, W.H.K., Kanamori, H., Jennings, P.C. & Kisslinger, C., Academic Press.

Myers, S.C., Johannesson, G. & Hanley, W., 2007. A Bayesian hierarchical method for multiple-event seismic location, *Geophys. J. Int.*, **171**(3), 1049–1063.

Myers, S.C., Johannesson, G. & Hanley, W., 2009. Incorporation of probabilistic seismic phase labels into a Bayesian multiple-event seismic locator, *Geophys. J. Int.*, **177**(1), 193–204.

Myers, S.C., Johannesson, G. & Simmons, N.A., 2011. Global-scale P wave tomography optimized for prediction of teleseismic and regional travel times for Middle East events: 1. Data set development, *J. Geophys. Res.*, **116**(B4), B04304, doi:10.1029/2010JB007967.

O'Hagan, A., 1994. *Kendall's Advanced Theory of Statistics, Vol. 2b: Bayesian Statistics*, John Wiley and Sons.

Taihoku Meteorological Observatory, 1936. Report of the severe Hsinchu-Taichung earthquake of April 21, 1935, pp. 160, Tech. Rep.

Tarantola, A. & Valette, B., 1982. Inverse problems = quest for information, *J. Geophys.*, **50**, 159–170.

Thurber, C.H., 2011. Earthquake, location techniques, in *Encyclopedia of Solid Earth Geophysics*, pp. 201–207, ed. Gupta, H.K., Springer-Verlag.

Wu, Y.-M., Shyu, J. B.H., Chang, C.-H., Zhao, L., Nakamura, M. & Hsu, S.-K., 2009. Improved seismic tomography offshore northeastern Taiwan: implications for subduction and collision processes between Taiwan and the southernmost Ryukyu, *Geophys. J. Int.*, **178**(2), 1042–1054.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

A set of text files are provided that describe the posterior PDF for the source location of the 1909 Taipei earthquake using either a Laplacian likelihood function or a Gaussian likelihood function. Each file has 3 lines corresponding to latitude (in degrees), longitude (in degrees), and the natural log of source depth (in kilometres).

laplacian posterior mean.txt Sample mean of the posterior PDF computed with a Laplacian likelihood function: $\bar{\theta} = \frac{1}{N}\sum_{i=1}^{N}\theta_i$.

laplacian central moment2.txt Second central moment of the posterior PDF computed with a Laplacian likelihood function: $\frac{1}{N}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^2$.

laplacian central moment3.txt Third central moment of the posterior PDF computed with a Laplacian likelihood function: $\frac{1}{N}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^3$.

laplacian central moment4.txt Fourth central moment of the posterior PDF computed with a Laplacian likelihood function: $\frac{1}{N}\sum_{i=1}^{N}(\theta_i - \bar{\theta})^4$.

laplacian posterior correlation.txt Correlation matrix for the posterior PDF computed with a Laplacian likelihood function. This file contains a 3×3 correlation matrix whose rows and columns correspond to latitude, longitude, and ln source depth, in that order.

gaussian posterior mean.txt Same as laplacian posterior mean.txt for the posterior PDF computed with a Gaussian likelihood function.

gaussian central moment2.txt Same as laplacian central moment2.txt for the posterior PDF computed with a Gaussian likelihood function.

gaussian central moment3.txt Same as laplacian central moment3.txt for the posterior PDF computed with a Gaussian likelihood function.

gaussian central moment4.txt Same as laplacian central moment4.txt for the posterior PDF computed with a Gaussian likelihood function.

gaussian posterior correlation.txt Same as laplacian posterior correlation.txt for the posterior PDF computed with a Gaussian likelihood function.

(http://gji.oxfordjournals.org/lookup/suppl/doi:10.1093/gji/ggu201/-/DC1)

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.